A Mixture-Model Approach to the Analysis of Survival Data

Lynn Kuo
Fengchun Peng

Abstract. We study a mixture model for survival data where covariates may influence both the incidence probabilities and their conditional latency distributions. The data may include the exactly observed, the right-censored, and the interval-censored failure times. We apply the EM algorithm to find the maximum likelihood estimate. We also carry out a Markov chain Monte Carlo algorithm for Bayesian inference. Model selection methods based on the predictive density for cross-validated data are developed. These methods allow us to assess whether simpler models would suffice as opposed to the mixture models. The potential of the methods is illustrated with the flour beetle data given by Hewlett (1974).

1. Introduction

In some analyses of failure-time data, it has been observed that a group of subjects may not react to treatments. A mixture model that incorporates different latency distributions for different groups seems to be appropriate. In this chapter we consider a mixture model where covariates may influence both the incidence probabilities and the conditional latency distributions. Let $\mathbf{x}$ denote the covariate vector $(1 \times q)$ associated with a subject of lifetime T. Then the mixture model assumes that the density of T is

$$f(t|\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{J} p_j(\mathbf{x}, \boldsymbol{\rho}) f_j(t|\mathbf{x}, \boldsymbol{\phi}_j), \tag{1.1}$$

where $\sum_{j=1}^{J} p_j(\mathbf{x}, \boldsymbol{\rho}) = 1$. Let Y be an index variable for the subpopulations. We use $p_j(\mathbf{x}, \boldsymbol{\rho})$ to denote the mixing probability, $P(Y = j|\mathbf{x}, \boldsymbol{\rho})$, also called the incidence probability for the j^{th} subpopulation. We use $f_j(t|\mathbf{x}, \boldsymbol{\phi}_j)$ to denote $f_j(t|Y = j, \mathbf{x}, \boldsymbol{\phi}_j)$, the conditional probability density (conditional latency density) function of the failure time for the j^{th} subpopulation. We assume f_j is a continuous density and indexed by an unknown parameter $\boldsymbol{\phi}_j$ that can be a vector. Moreover, we use $\boldsymbol{\theta} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_J, \boldsymbol{\rho})$ to denote the collection of all unknown parameters. The covariate $\mathbf{x}$ may include 1, the dosage (or log-dose) level, and other explanatory variables. In this study we only deal with time-independent explanatory variables.

Logistic regression links can be chosen for the incidence probabilities. For example, we can choose

$$p_j(\mathbf{x}, \boldsymbol{\rho}) = \frac{e^{\xi_j}}{\sum_{l=1}^{J} e^{\xi_l}}, \tag{1.2}$$

where $\xi_j = \mathbf{x}\boldsymbol{\rho}_j^T$, with $\boldsymbol{\rho}_j^T$ to be the transpose of $\boldsymbol{\rho}_j$ $(1 \times q)$ and $\boldsymbol{\rho} = (\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_J)$. In addition to the logistic link, other models such as the normalized probit link or the normalized complementary log-log link (McCullagh and Nelder, 1989) can be considered for the incidence probabilities.

The cumulative distribution function for T is

$$F(t|\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{J} p_j(\mathbf{x}, \boldsymbol{\rho}) F_j(t|\mathbf{x}, \boldsymbol{\phi}_j), \tag{1.3}$$

where F_j is the distribution function for f_j.

This mixture model has been studied by Farewell (1982), Larson and Dinse (1985), Pack and Morgan (1990), Boos and Brownie (1991), Kuk and Chen (1992), and Taylor (1995). The paper of Boos and Brownie considers the failure times to be independent and identically distributed (i.i.d.) as in (1.1). The rest of the papers include censored failure times. Pack and Morgan consider interval-censored and right-censored time-to-response data for quantal assay. Larson and Dinse and Kuk and Chen consider proportional hazards models for the conditional latency distributions. Taylor considers Kaplan-Meier type formulations for the conditional latency distributions.

In this chapter, we consider data that include both right-censored and interval-censored observations. Our methodologies can be applied to left-censored observations as well. We provide two methodologies for inference. One is an EM algorithm or a Monte Carlo EM algorithm (MCEM) to obtain the maximum likelihood estimate of $\boldsymbol{\theta}$. The other is a Markov chain Monte Carlo (MCMC) algorithm for Bayesian inference.

To apply the EM algorithm, we need to consider a complete likelihood of a product of components for i.i.d. observations as opposed to the mixture likelihood with censored observations. This is done by augmenting the original data with two classes of latent variables. One is the truncated random variable denoted by $\mathbf{w}$ that allows us to consider a likelihood for i.i.d. observations without censoring. The other is the index variable $\mathbf{z}$ that converts the mixture model to a model of independent components. Although the E-step with respect to the latent variables $\mathbf{w}$ is often difficult to evaluate, this is not the case for $\mathbf{z}$ when it is conditioned on $\mathbf{w}$. We employ a MCEM algorithm (Wei and Tanner, 1990) to approximate the E-step, where integration with respect to $\mathbf{w}$ is done by Monte Carlo integration and integration with respect to $\mathbf{z}$ given $\mathbf{w}$ is evaluated exactly. Then we apply a maximization step to update the parameters until convergence.

The MCMC algorithm consists of generating the two latent variables $\mathbf{w}$, $\mathbf{z}$ and the unknown parameter $\boldsymbol{\theta}$ iteratively. Let $\mathbf{t}$ denote the data set that contains censored observations. Starting at the initial choices of the variates, $\mathbf{w}^{(0)}$, $\mathbf{z}^{(0)}$, and $\boldsymbol{\theta}^{(0)}$, sample $\mathbf{w}^{(1)}$ from $f(\mathbf{w}|\mathbf{z}^{(0)}, \boldsymbol{\theta}^{(0)}, \mathbf{t})$; sample $\mathbf{z}^{(1)}$ from $f(\mathbf{z}|\mathbf{w}^{(1)}, \boldsymbol{\theta}^{(0)}, \mathbf{t})$; and sample $\boldsymbol{\theta}^{(1)}$ from $f(\boldsymbol{\theta}|\mathbf{w}^{(1)}, \mathbf{z}^{(1)}, \mathbf{t})$. Continue iteration until convergence. The f function here is a generic name for the conditional densities.

Both the MCEM and the MCMC algorithms use data augmentation with latent variables to simplify the computation. The techniques used in this paper are related to that in many other papers. We only mention a few here. Dempster, Laird, and Rubin (1977) introduced the EM algorithm for analysis with aggregated data or with missing data. McLachlan and Jones (1988) developed the EM algorithm for grouped and truncated data. Wei and Tanner (1990) proposed MCEM when the E-step is difficult to derive. Jordan and Jacobs (1994) showed how the parameters of the mixtures-of-experts (as well as the hierarchical mixtures-of-experts) architecture can be estimated using the EM algorithm. The MCMC algorithm was introduced by Geman and Geman (1984) for image restoration and was generalized by Tanner and Wong (1987) for missing value models. Gelfand and Smith (1990) applied it to Bayesian analysis. Diebolt and Robert (1994) developed a Gibbs sampler for i.i.d. observations for the usual mixture model, that is a simpler version of (1.1) without the covariates. Dey, Kuo, and Sahu (1995) studied a predictive approach to selecting the number of components in the mixture model. We are extending the two papers to censored data. The basic principle in the extension can be seen in Gelfand, Smith, and Lee (1992) and Kuo and Smith (1992). Peng, Jacobs, and Tanner (1996) present a hierarchical mixtures-of-experts architecture using both EM and a Gibbs sampler in a classification problem.

The rest of the paper is organized as follows. Section 2 describes the likelihood. Section 3 describes the EM and MCEM algorithms. Section 4 describes the MCMC algorithm. Model selection based on the predictive approach is discussed in Section 5. An illustration using the flour beetle data given by Hewlett (1974) is given in Section 6.

2. Likelihood

We assume that failure times are independently distributed, also independent from the censoring process. For each individual we observe the time to failure, a right-censored, or an interval-censored failure time. That is, for the right-censored individual we only know that the time to failure is greater than the censored time, while for the interval-censored individual, the time to failure is only known to be between two time points.

The likelihood is then (Cox and Oakes, 1984)

$$L(\boldsymbol{\theta}|\mathbf{t}, \mathbf{X}) = \prod_{i \in \mathcal{U}} f(t_i|\mathbf{x}_i, \boldsymbol{\theta}) \prod_{i \in \mathcal{C}} \{1 - F(t_i^+|\mathbf{x}_i, \boldsymbol{\theta})\}$$

$$\times \prod_{i\in\mathcal{I}} \{F(t_{iU}|\mathbf{x}_i,\boldsymbol{\theta}) - F(t_{iL}|\mathbf{x}_i,\boldsymbol{\theta})\}, \tag{2.1}$$

where $\mathbf{t}$ denotes the data set with the i^{th} entry denoted by t_i if the i^{th} subject is exactly observed, by t_i^+ if it is right-censored at t_i, and by (t_{iL}, t_{iU}) if the i^{th} subject is interval-censored between t_{iL} and t_{iU}; the sets $\mathcal{U}$, $\mathcal{C}$ and $\mathcal{I}$ are the sets of indices for the exactly observed, the right-censored, and the interval-censored subjects, respectively. We assume there are N subjects in the study. Therefore, we use $\mathbf{X}$ to denote $(\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T)$, the matrix of all covariates.

3. EM and Monte Carlo EM

In this section, we develop a Monte Carlo EM algorithm to maximize $\boldsymbol{\theta}$ in the likelihood. First observe that part of the likelihood in (2.1) for the censored data is a product of *integrals* of the mixture density in (1.1). This incomplete expression makes maximization difficult even for simple latency distributions such as normal, gamma, etc. For each subject i, $i = 1, \ldots, N$, we consider the latent variable w_i that is truncated between (t_{iL}, t_{iU}) if it is interval-censored, truncated at t_i^+ if it is right-censored, and is t_i if it is exactly observed. Then the joint density of the data and the latent variables w's has i.i.d. components, each distributed as (1.1). This will alleviate the difficulties in maximization with incomplete likelihood as discussed earlier. More specifically, if the i^{th} subject is interval-censored between (t_{iL}, t_{iU}), then we can generate a latent variable w_i from the truncated density $f(w_i)/[F(t_{iU}) - F(t_{iL})]\mathrm{I}(t_{iL} < w_i < t_{iU})$, where f is given in (1.1). This can be done by setting

$$w_i = F^{-1}\{F(t_{iL}|\mathbf{x}_i,\boldsymbol{\theta}) + U[F(t_{iU}|\mathbf{x}_i,\boldsymbol{\theta}) - F(t_{iL}|\mathbf{x}_i,\boldsymbol{\theta})]\}, \tag{3.1}$$

where $U \sim$ uniform$(0,1)$ and F^{-1} is the inverse function of F. Similarly, if the i^{th} subject is right-censored at t_i^+, then we can generate a latent variable w_i from the truncated density $f(w_i)/[1 - F(t_i^+)]\mathrm{I}(t_i^+ < w_i < \infty)$ by setting

$$w_i = F^{-1}\{F(t_i^+|\mathbf{x}_i,\boldsymbol{\theta}) + U[1 - F(t_i^+|\mathbf{x}_i,\boldsymbol{\theta})]\}. \tag{3.2}$$

More details about Equations (3.1) and (3.2) are given by Devroye (1986, pp. 38-39). Note from (3.1) that w_i can be generated as the unique root of $F(w_i) = F(t_{iL}|\mathbf{x}_i,\boldsymbol{\theta}) + U[F(t_{iU}|\mathbf{x}_i,\boldsymbol{\theta}) - F(t_{iL}|\mathbf{x}_i,\boldsymbol{\theta})]$ and similarly for (3.2). They can be easily implemented using the IMSL subroutine "ZREAL" in FORTRAN. If the i^{th} subject is exactly observed at t_i, then we set $w_i = t_i$.

For each i, we generate w_i independently as discussed above. Let $\mathbf{w}$ denote $(w_1, \ldots, w_N)$. Then the likelihood of $\boldsymbol{\theta}$ given $\mathbf{w}, \mathbf{t}, \mathbf{X}$ is

$$L(\boldsymbol{\theta}|\mathbf{w}, \mathbf{t}, \mathbf{X}) = \prod_{i=1}^{N} f(w_i \mid \mathbf{x}_i, \boldsymbol{\theta}) = \prod_{i=1}^{N} \left(\sum_{j=1}^{J} p_j(\mathbf{x}_i, \boldsymbol{\rho}) f_j(w_i \mid \mathbf{x}_i, \boldsymbol{\phi}_j) \right). \tag{3.3}$$

Next, we make use of the fact that a mixture model can always be expressed by a product of its components. This is done by considering another latent variable $\mathbf{z}_i = (z_{i1}, \ldots, z_{iJ})$, where $z_{ij} = 1$ if w_i is considered as coming from the subpopulation having pdf f_j and $\sum_{j=1}^{J} z_{ij} = 1$. The vector $\mathbf{z}_i$ follows a multinomial distribution, i.e.

$$\mathbf{z}_i|\boldsymbol{\theta}, \mathbf{w}, \mathbf{t}, \mathbf{X} \sim Mult(1, h_{i1}, \ldots, h_{iJ}), \tag{3.4}$$

with h_{ij}, the abbreviation for $h_{ij}(\boldsymbol{\theta}|w_i, \mathbf{t}, \mathbf{X})$, given by

$$h_{ij}(\boldsymbol{\theta}|w_i, \mathbf{t}, \mathbf{X}) = \frac{p_j(\mathbf{x}_i, \boldsymbol{\rho}) f_j(w_i|\mathbf{x}_i, \boldsymbol{\phi}_j)}{\sum_j p_j(\mathbf{x}_i, \boldsymbol{\rho}) f_j(w_i|\mathbf{x}_i, \boldsymbol{\phi}_j)}. \tag{3.5}$$

Let $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$. Then the augmented likelihood for the completely imputed data $\mathbf{z}$ and $\mathbf{w}$ is

$$L(\boldsymbol{\theta}|\mathbf{z}, \mathbf{w}, \mathbf{t}, \mathbf{X}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \{p_j(\mathbf{x}_i, \boldsymbol{\rho}) f_j(w_i|\mathbf{x}_i, \boldsymbol{\phi}_j)\}^{z_{ij}}. \tag{3.6}$$

The complexity of fitting the mixture model is greatly simplified by working with the augmented likelihood because we only need to evaluate $p_j(\mathbf{x}_i, \boldsymbol{\rho}) \times f_j(w_i|\mathbf{x}_i, \boldsymbol{\phi}_j)$ when $z_{ij} = 1$. Suppose the current EM iteration is at the k^{th} stage. The E-step in the E-M algorithm consists of evaluating

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= \int \log[L(\boldsymbol{\theta}|\mathbf{z}, \mathbf{w}, \mathbf{t}, \mathbf{X})] f(\mathbf{z}|\mathbf{w}, \boldsymbol{\theta}^{(k)}, \mathbf{t}, \mathbf{X}) f(\mathbf{w}|\boldsymbol{\theta}^{(k)}, \mathbf{t}, \mathbf{X}) d\mathbf{z} d\mathbf{w} \\ &= \sum_{i=1}^{N} \sum_{j=1}^{J} E\{h_{ij}(\boldsymbol{\theta}^{(k)}|w_i, \mathbf{t}, \mathbf{X}) \log[p_j(\mathbf{x}_i, \boldsymbol{\rho}) f_j(w_i|\mathbf{x}_i, \boldsymbol{\phi}_j)]\} \end{aligned} \tag{3.7}$$

where the expectation on the right-hand side of the second equality is with respect to the truncated random variable $\mathbf{w}$. The M-step requires maximizing the above expression as a function of $\boldsymbol{\theta}$. The maximization for $\boldsymbol{\phi}_j$ and $\boldsymbol{\rho}$ can be done separately. For many standard probability distributions, statistical packages are available to handle the maximization. We continue this iteration using (3.7) until the two consecutive $\boldsymbol{\theta}$s are close enough. Mixture models are usually multi-modal; hence the EM solutions are not unique and depend on the starting values. Therefore exploration with different starting points is needed to obtain the global maximum.

Now we extend the EM to the MCEM algorithm. When the expectation is hard to evaluate with respect to the truncated random variables $\mathbf{w}$, we circumvent it by multiple imputation by generating M copies of $\mathbf{w}$, where the m^{th} copy of $\mathbf{w}$ denoted by $\mathbf{w}^m$ equals $(w_1^m, \ldots, w_N^m)$. Then for each $\mathbf{w}^m$, we generate $\mathbf{z}^m = (\mathbf{z}_1^m, \ldots, \mathbf{z}_N^m)$, where $\mathbf{z}_i^m$ is generated by (3.4) with $h_{ij}^m(\boldsymbol{\theta}^{(k)}|w_i^m, \mathbf{t}, \mathbf{X})$ defined in (3.5), where w_i is replaced by w_i^m and $\boldsymbol{\theta}$ is replaced by $\boldsymbol{\theta}^{(k)}$. Therefore the E-step in (3.7) can be approximated by

$$\hat{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{j=1}^{J} h_{ij}^m(\boldsymbol{\theta}^{(k)}|w_i^m, \mathbf{t}, \mathbf{X}) \{\log[p_j(\mathbf{x}_i, \boldsymbol{\rho}) f_j(w_i^m|\mathbf{x}_i, \boldsymbol{\phi}_j)]\}.$$

4. Gibbs Sampler

Our next objective is a general Bayesian approach to mixture models for the survival data with censoring. Our sampling-based approach enables us to construct credible regions for parameters or functionals of interest for any sample size. Moreover, by plotting the entire posterior distribution of $\boldsymbol{\theta}$, we can detect interesting features of the posterior distribution, such as skewness, heavy tailedness, etc. Let $\pi(\boldsymbol{\theta})$ denote the prior density for $\boldsymbol{\theta}$; the posterior density is proportional to $L(\boldsymbol{\theta}|\mathbf{t}, \mathbf{X}) \times \pi(\boldsymbol{\theta})$. We first consider the data augmentation idea of the latent variables $\mathbf{z}$ discussed earlier, so updating on the $\boldsymbol{\theta}$ can be made simple in the MCMC by means of considering a product of the components. Now we extend these techniques to handle censored data. In addition to $\mathbf{z}$, we augment the data with another class of latent variables $\mathbf{w}$ that treats the unobserved failure times due to censoring as missing. This enables us to consider the augmented likelihood as given in (3.6). Then simulating $\boldsymbol{\theta}$ given $\mathbf{w}, \mathbf{z}, \mathbf{t}$, and $\mathbf{X}$ can be performed either from standard distributions or by using the Metropolis algorithm (1953).

Now let us briefly discuss the Gibbs sampler. Start with some initial values of $\boldsymbol{\theta}^{(0)}, \mathbf{z}^{(0)}, \mathbf{w}^{(0)}$. One of the values could be the EM or the MCEM solution of the mixture model. We can also perturb it with small errors for the starting points for replications of the Markov chain. We iteratively generate $\mathbf{w}$, $\mathbf{z}$, and $\boldsymbol{\theta}$ from the following full conditional distributions:
step 1: Given $\boldsymbol{\theta}^{(0)}, \mathbf{t}, \mathbf{X}$, we generate $w_i^{(1)}$ by (3.1) if the i^{th} subject is interval censored; generate $w_i^{(1)}$ by (3.2), if it is right censored; set it to t_i, if it is exactly observed. We do this independently for $i = 1, \ldots, N$.
step 2: Given $\mathbf{w}^{(1)}$, $\boldsymbol{\theta}^{(0)}, \mathbf{t}, \mathbf{X}$, we generate $\mathbf{z}_i^{(1)} \sim Mult(1, h_{i1}, \ldots, h_{iJ})$, where $h_{ij} = h_{ij}(\boldsymbol{\theta}^{(0)}|w_i^{(1)}, \mathbf{t}, \mathbf{X})$ as given in (3.5). This is done for all $i = 1, \ldots, N$.
step 3: Given $\mathbf{z}^{(1)}, \mathbf{w}^{(1)}, \mathbf{t}, \mathbf{X}$, we generate $\boldsymbol{\theta}^{(1)}$ from $p(\boldsymbol{\theta}|\mathbf{w}^{(1)}, \mathbf{z}^{(1)}, \mathbf{t}, \mathbf{X}) \propto \pi(\boldsymbol{\theta}) L(\boldsymbol{\theta}|\mathbf{z}^{(1)}, \mathbf{w}^{(1)}, \mathbf{t}, \mathbf{X})$, where $L(\boldsymbol{\theta}|\mathbf{z}, \mathbf{w}, \mathbf{t}, \mathbf{X})$ is given in (3.6).
Then we continue to iterate by repeating steps 1-3 where the superscripts

in the stage indicators are incremented. Convergence can be monitored by the Gelman and Rubin (1992) method.

5. Model Selection

Given the mixture model in (1.1), we need to know what is the best number of mixtures to fit the data. We also have many choices for the mixing probabilities that include the logistic link and the probit link, and many choices for the latency distributions that include normal, log-normal, exponential, and Weibull distributions. We explore a predictive approach for model selection.

The basic idea is simple. We use part of the data to fit the model and the remaining part to test it. We divide our data into G groups of roughly equal size. For the data in the g^{th} group, we evaluate the predictability of a model by evaluating the predictive likelihood of the g^{th} group relative to the posterior distribution of $\boldsymbol{\theta}$ given the rest of the data. Let $\mathbf{t}_g$ denote the data set for the g^{th} group that may include censored data. Similarly, we use $\mathbf{X}_g$ to denote the covariates for the g^{th} group. Then the remaining data are denoted by $\mathbf{t}_{-g}$ and $\mathbf{X}_{-g}$. The conditional predictive ordinate for the g^{th} group (GCPO) is defined by

$$\begin{aligned} p(\mathbf{t}_g|\mathbf{t}_{-g},\mathbf{X}_{-g}) &= \int p(\mathbf{t}_g|\boldsymbol{\theta},\mathbf{X}_g)\pi(\boldsymbol{\theta}|\mathbf{t}_{-g},\mathbf{X}_{-g})d\boldsymbol{\theta} \\ &= \int L(\boldsymbol{\theta}|\mathbf{t}_g,\mathbf{X}_g)\pi(\boldsymbol{\theta}|\mathbf{t}_{-g},\mathbf{X}_{-g})d\boldsymbol{\theta}. \end{aligned} \tag{5.1}$$

This GCPO assesses the predictability of our model for the new data $\mathbf{t}_g$. We do this for all g, $g = 1, \ldots, G$. The pseudo-marginal predictive likelihood for the whole data is defined to be the product of these GCPO's,

$$PGCPO_{(\mathbf{t},\mathbf{X})} = \prod_{g=1}^{G} p(\mathbf{t}_g|\mathbf{t}_{-g},\mathbf{X}_{-g}). \tag{5.2}$$

The predictive pseudo-likelihood depends on the model. We use $PGCPO_{(\mathbf{t},\mathbf{X})}(M)$ to denote the predictive pseudo-likelihood for the model M. Therefore, we select the best model that has the largest $P\widehat{GCPO}_{(\mathbf{t},\mathbf{X})}(M)$ among the class of models $\{M \in \mathcal{M}\}$.

Usually, if the data size is large, a G of two may be sufficient. (cf. p. 240 of Efron and Tibshirani, 1993, for a related problem on estimating the prediction error). If the data size is small, we may need more groups. If G is N, then the method is just leave-one-out cross-validation. Then (5.2) reduces to the usual pseudo-marginal likelihood as defined in Gelfand, Dey, and Chang (1992).

On the computation for (5.2), if G is small, we might use a brute-force Monte Carlo integration. For the g^{th} group, we apply the Gibbs sampler in Section 4 using the new data set with the g^{th} group deleted. Then we apply Monte Carlo integration to (5.1) by averaging $L(\boldsymbol{\theta}|\mathbf{t}_g, \mathbf{X}_g)$ over the $\boldsymbol{\theta}$ sampled from the Gibbs sampler applied to the new data set $(\mathbf{t}_{-g}, \mathbf{X}_{-g})$. If $G = N$, it would be too computing intensive to repeat the Gibbs sampler N times, every time using a different data set. Then a harmonic mean estimator of the CPO (GCPO in (5.1 with $G = N$)) can be obtained from the Gibbs sample based on the full data. Let $\mathcal{K}$ denote an index set that counts the usable variates $\boldsymbol{\theta}^{(k)}$ generated by the Gibbs sampler applied to the full data as in Section 4. These variates can be a second half of a single chain, or a second half of all multiple chains. Let K denote the size of $\mathcal{K}$. Then

$$\hat{p}(t_i|\mathbf{t}_{-i}, \mathbf{X}_{-i}) = \begin{cases} K\{\sum_{k\in\mathcal{K}} f(t_i|\mathbf{x}_i, \boldsymbol{\theta}^{(k)})]^{-1}\}^{-1} & \text{for } i \in \mathcal{U}; \\ K\{\sum_{k\in\mathcal{K}}[1 - F(t_i^+|\mathbf{x}_i, \boldsymbol{\theta}^{(k)})]^{-1}\}^{-1} & \text{for } i \in \mathcal{C}; \\ K\{\sum_{k\in\mathcal{K}}[F(t_{iU}|\mathbf{x}_i, \boldsymbol{\theta}^{(k)}) - F(t_{iL}|\mathbf{x}_i, \boldsymbol{\theta}^{(k)})]^{-1}\}^{-1} & \text{for } i \in \mathcal{I}. \end{cases}$$

6. Example

As an illustration, consider the data set in Table 1 given in Hewlett (1974). It was also used by Diggle and Gratton (1984), Pack and Morgon (1990), and several other authors. The data set consists of 317 male adult flour beetles that were exposed separately to pyrethrum, a plant-based insecticide. Among the 317 males, 144, 69, 54, and 50 males were sprayed of pyrethrum at concentrations of 0.20, 0.32, 0.50, 0.80 mg/cm^2, respectively. The equivalent log concentrations (denoted by log-doses) are -1.61, -1.14, -0.69, and -0.22. Similarly, Hewlett (1974) and Pack and Morgan (1990, p.750) contain a data set for the females, the group sizes corresponding to the four dosages are 152, 81, 44, 47.

Let $x_c = \log(d_c)$, where d_c is the dosage given to the beetles in the c^{th} column. The survival time of a beetle given dosage x_c is modeled by a mixture of normal and exponential densities:

$$f(t|x_c, \boldsymbol{\theta}) = p_1(x_c, \gamma, \tau) f_1(t|x_c, \alpha, \beta, \sigma) + p_2(x_c, \gamma, \tau) f_2(t|\lambda); \quad (6.3)$$

where

$$\begin{aligned} p_1(x_c, \gamma, \tau) &= \frac{e^{\gamma+\tau x_c}}{1 + e^{\gamma+\tau x_c}}; \\ f_1(t|x_c, \alpha, \beta, \sigma) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(t - \alpha - \beta x_c)^2}{2\sigma^2}\right\}; \\ f_2(t|\lambda) &= \lambda \exp\{-\lambda t\} \mathrm{I}\{t > 0\}; \text{ and} \\ p_2(x_c, \gamma, \tau) &= 1 - p_1(x_c, \gamma, \tau). \end{aligned}$$

This model suggests a normal distribution for the failure times in the susceptible group with the mean to be a linear function of the log dose as in Boos and Brownie (1991). Because the other group is assumed to be more resistant to treatment, we choose an exponential distribution independent of the dosages. Other selections of standard distributions that might fit the data as well will be explored later. Because the mixing probabilities are directly related to the dosages, we consider logistic regression links for the incidence probabilities. Pack and Morgan (1990) ignore the contribution from the subpopulation of the long-term survivors for $t \leq 13$ in their numerical examples. Our model includes contributions from both the susceptible and the long-term survivor groups over the entire range of the survival times. These contributions are weighted by the mixing probabilities that correspond to the proportions of the subgroups. These two aspects are the advantages of using a mixture model.

Table 1: Male Flour Beetle (*Tribolium castaneum*) Data. The numbers indicate the number dead per day, the row 14 gives the number survived after day 13.

r/c	Log-dose(mg/cm^2)			
	-1.61	-1.14	-0.69	-0.22
1	3	7	5	4
2	11	10	8	10
3	10	11	11	8
4	7	16	15	14
5	4	3	4	8
6	3	2	2	2
7	2	1	1	1
8	1	0	1	0
9	0	0	0	0
10	0	0	0	1
11	0	0	0	0
12	1	0	0	0
13	1	0	0	0
14	101	19	7	2
Total	144	69	54	50

6.1 *EM Algorithm for the Specific Example*

Now we derive the EM algorithm for this example. Let n_{rc} denote the cell count in the r^{th} row and the c^{th} column, where $c = 1, \ldots, 4$ is the index for the dose levels and $r = 1, \ldots, 14$ is the index on the days. For simplicity, we write $\sum_{rc}$ for $\sum_{c=1}^{4} \sum_{r=1}^{14}$. We treat the data in the r^{th} row as interval censored for $r \leq 13$, that is, beetles in the r^{th} row die between the $r - 1^{th}$ and r^{th} days; and data in the 14^{th} row as right-censored with survival until at least the 13^{th} day.

For this example, we are able to evaluate the E-step in closed form. Therefore, no MCEM is needed. A related EM algorithm can be found in

McLachlan and Jones (1988). Suppose the current EM iteration is at the k^{th} stage. Let $\boldsymbol{\theta}^{(k)} = \left(\alpha^{(k)}, \beta^{(k)}, (\sigma^{(k)})^2, \lambda^{(k)}, \gamma^{(k)}, \tau^{(k)}\right) = \left(\boldsymbol{\phi}_1^{(k)}, \boldsymbol{\phi}_2^{(k)}, \boldsymbol{\rho}_1^{(k)}\right)$, where $\boldsymbol{\phi}_1$ denotes $(\alpha, \beta, \sigma^2)$, $\boldsymbol{\phi}_2$ denotes λ, and $\boldsymbol{\rho}_1$ denotes (γ, τ). Note we set $\boldsymbol{\rho}_2 = (0,0)$. For the E-step, we need to evaluate the expectation in (3.7) with respect to the truncated random variable $\mathbf{w}$. Because of the grouped data, the summation over i in (3.7) can be written as $\sum_{rc}$ with w_i replaced by W_{rc}, where W_{rc} is distributed as a truncated version of (6.3) with the current value of $\boldsymbol{\theta}^{(k)}$. We use (a_{rc}, b_{rc}) to denote the range of W_{rc}. We write $p_j^{(k)}$ for $p_j^{(k)}(x_c, \boldsymbol{\rho})$ for short. For $j = 1$ or 2, let

$$h_j(W_{rc}; \boldsymbol{\theta}^{(k)}) = \frac{p_j^{(k)} f_j^{(k)}(W_{rc}|x_c, \boldsymbol{\phi}_j^{(k)})}{\sum_j p_j^{(k)} f_j^{(k)}(W_{rc}|x_c, \boldsymbol{\phi}_j^{(k)})}.$$

Therefore, for the E-step, we need to evaluate the following expectations:

$$A_{s,jrc}^{(k)} = E^{(k)}[h_j(W_{rc}; \boldsymbol{\theta}^{(k)}) W_{rc}^s],$$

for $s = 0, 1$ and

$$A_{2,1rc}^{(k)} = E^{(k)}[h_1(W_{rc}; \boldsymbol{\theta}^{(k)})(W_{rc} - \mu_1^{(k+1)})^2].$$

For simplicity, we write μ_1, instead of $\mu_1(x_c)$, for $\alpha + \beta x_c$. It is straightforward to verify the following equalities for $c = 1, \ldots, 4$ and $r = 1, \ldots, 14$, where $b_{14c} = \infty$, for all c:

$$A_{s,jrc}^{(k)} = \frac{p_j^{(k)} G_{s,jrc}^{(k)}}{F^{(k)}(b_{rc}) - F^{(k)}(a_{rc})}, \ (s = 0, 1, 2),$$

where

$$\begin{aligned}
G_{0,jrc}^{(k)} &= H_{0,jrc}^{(k)}, \\
G_{1,1rc}^{(k)} &= \mu_1^{(k)} H_{0,1rc}^{(k)} - (\sigma^{(k)})^2 H_{1,1rc}^{(k)}, \\
G_{1,2rc}^{(k)} &= \exp\{-\lambda^{(k)} a_{rc}\}(a_{rc} + 1/\lambda^{(k)}) - \exp\{-\lambda^{(k)} b_{rc}\}(b_{rc} + 1/\lambda^{(k)}), \\
G_{2,1rc}^{(k)} &= (\sigma^{(k)})^2 [H_{0,1rc}^{(k)} + (2\mu_1^{(k+1)} - \mu_1^{(k)}) H_{1,1rc}^{(k)} - H_{2,1rc}^{(k)}] + (\mu_1^{(k+1)} - \mu_1^{(k)})^2 H_{0,1rc}^{(k)},
\end{aligned}$$

and where

$$\begin{aligned}
H_{0,jrc}^{(k)} &= F_j^{(k)}(b_{rc}) - F_j^{(k)}(a_{rc}), \\
H_{1,jrc}^{(k)} &= f_j^{(k)}(b_{rc}) - f_j^{(k)}(a_{rc}), \\
H_{2,1rc}^{(k)} &= b_{rc} f_1^{(k)}(b_{rc}) - a_{rc} f_1^{(k)}(a_{rc}).
\end{aligned}$$

Therefore, we obtain the following equations for the M-step:

$$\alpha^{(k+1)} = \frac{\sum_{rc} n_{rc} A_{1,1rc}^{(k)} - \beta^{(k+1)} \sum_{rc} n_{rc} A_{0,1rc}^{(k)} x_c}{\sum_{rc} n_{rc} A_{0,1rc}^{(k)}},$$

$$\beta^{(k+1)} = \frac{\sum_{rc} n_{rc} A_{1,1rc}^{(k)} x_c \sum_{rc} n_{rc} A_{0,1rc}^{(k)} - \sum_{rc} n_{rc} A_{1,1rc}^{(k)} \sum_{rc} n_{rc} A_{0,1rc}^{(k)} x_c}{\sum_{rc} n_{rc} A_{0,1rc}^{(k)} x_c{}^2 \sum_{rc} n_{rc} A_{0,1rc}^{(k)} - (\sum_{rc} n_{rc} A_{0,1rc}^{(k)} x_c)^2},$$

$$\left(\sigma^{(k+1)}\right)^2 = \frac{\sum_{rc} n_{rc} A_{2,1rc}^{(k)}}{\sum_{rc} n_{rc} A_{0,1rc}^{(k)}},$$

$$\lambda^{(k+1)} = \frac{\sum_{rc} n_{rc} A_{0,2rc}^{(k)}}{\sum_{rc} n_{rc} A_{1,2rc}^{(k)}},$$

and the M-step for γ and τ is computed using the Newton-Raphson method (Boos and Brownie, 1991, and Jordan and Xu, 1993). Let

$$\boldsymbol{\rho}_1^{T(k+1)} = \boldsymbol{\rho}_1^{T(k)} + g\left(R^{(k)}\right)^{-1} \mathbf{e}^{(k)},$$

where the gradient vector at the k^{th} iteration is

$$\mathbf{e}^{(k)} = \sum_{rc} n_{rc} \left(E^{(k)}[h_1(W_{rc}; \boldsymbol{\theta}^{(k)})] - p_1^{(k)} \right) \frac{\partial \xi_1}{\partial \boldsymbol{\rho}_1^{T(k)}},$$

where $\xi_1 = \mathbf{x}\boldsymbol{\rho}_1^T$ and $\frac{\partial \xi_1}{\partial \boldsymbol{\rho}_1^{T(k)}} = \{1, x_c\}^T$. The Hessian matrix at the k^{th} iteration is

$$R^{(k)} = \sum_{rc} n_{rc} p_1^{(k)} (1 - p_1^{(k)}) \frac{\partial \xi_1}{\partial \boldsymbol{\rho}_1^{T(k)}} \frac{\partial \xi_1}{\partial \boldsymbol{\rho}_1^{(k)}},$$

where g is a learning rate.

6.2 Gibbs Samplers for the Specific Example

In the prior specification, we assume the components of $\boldsymbol{\theta}$ are independent. Flat priors are chosen for α, β, λ, and σ^2 because we lack precise knowledge about these parameters. Although the priors are improper, they result in proper posteriors in most situations. We use normal priors for γ and τ, both with mean zero and a large known variance σ_0^2, that is, $\gamma \sim N(0, \sigma_0^2)$ and $\tau \sim N(0, \sigma_0^2)$. Proper priors are chosen here due to the concern about the convergence of the posterior sample if flat priors were used for the parameters in the mixing probability (Diebold and Robert, 1994, p. 367).

We choose the covariate vector $\mathbf{x} = (1, x)$ where $x = \log(\text{dose level})$. As in the EM method, we treat each entry of Table 1 as interval censored except that the row labeled 14 is treated as right-censored. Therefore, reading from the first column, the data set t_1 to t_{144} consists of $\{\underbrace{(0,1), \ldots, (0,1)}_{3\ copies}, \underbrace{(1,2), \ldots, (1,2)}_{11\ copies}, \ldots, (12,13), \underbrace{13^+, \ldots, 13^+}_{101\ copies}\}$ with the same covariate $\mathbf{x}_i = (1, x_i) = (1, \log .2) = (1, -1.61)$ for $i = 1, \ldots, 144$. Similarly, we can list the rest of the data for the remaining dosages.

On implementing the Gibbs sampler, we follow steps 1–3 in Section 4. We first follow steps 1-2 to generate latent variable for w_i and $\mathbf{z}_i = (z_{i1}, z_{i2})$ for each i. Then we follow step 3 to update the parameters. Now we give more details for step 3, where the joint conditional density of $\boldsymbol{\theta}$ can be implemented sequentially by updating one variable at a time. Because given $\mathbf{w}$ and $\mathbf{z}$ (we omit "given $\mathbf{t}, \mathbf{x}$" for simpler notation), the parameters $\boldsymbol{\rho}_1 = (\gamma, \tau)$, $\boldsymbol{\phi}_1 = (\alpha, \beta, \sigma^2)$, and $\boldsymbol{\phi}_2 = \lambda$ are independent, we have some simplifications in generating the inner loop. That is, we generate $\alpha^{(1)}$ given $\beta^{(0)}, \sigma^{(0)}, \mathbf{w}^{(0)}, \mathbf{z}^{(0)}$; generate $\beta^{(1)}$ given $\alpha^{(1)}, \sigma^{(0)}, \mathbf{w}^{(0)}, \mathbf{z}^{(0)}$; generate ${\sigma^2}^{(1)}$ given $\alpha^{(1)}, \beta^{(1)}, \mathbf{w}^{(0)}, \mathbf{z}^{(0)}$; generate $\lambda^{(1)}$ given $\mathbf{w}^{(0)}$ and $\mathbf{z}^{(0)}$; and independently generate $\gamma^{(1)}$ and $\tau^{(1)}$ given $\mathbf{w}^{(0)}$ and $\mathbf{z}^{(0)}$. Then continue back to step 1 in Section 4 of the iteration. Let $z_{+1} = \sum_{i=1}^N z_{i1}$. Now we just list the conditional densities used in the inner loop of step 3.

$$\alpha | \ldots \sim N\left(\frac{\sum_{i=1}^N z_{i1}(w_i - \beta x_i)}{z_{+1}}, \ \frac{\sigma^2}{z_{+1}}\right); \tag{6.4}$$

$$\beta | \ldots \sim N\left(\frac{\sum_{i=1}^N z_{i1} x_i (w_i - \alpha)}{\sum_{i=1}^N z_{i1} {x_i}^2}, \ \frac{\sigma^2}{\sum_{i=1}^N z_{i1} {x_i}^2}\right); \tag{6.5}$$

$$\sigma^2 | \ldots \sim IG\left(\frac{z_{+1}}{2}, \ \frac{2}{C}\right), \tag{6.6}$$

where IG denotes the inverse gamma distribution with mean $C/(z_{+1} - 2)$, where $C = \sum_{i=1}^N z_{i1}(w_i - \alpha - \beta x_i)^2$;

$$\lambda | \ldots \sim \Gamma\left(N - z_{+1} + 1, \ \frac{1}{\sum_{i=1}^N w_i (1 - z_{i1})}\right), \tag{6.7}$$

where Γ denotes the gamma distribution with mean $(N - z_{+1} + 1)/(\sum_{i=1}^N w_i (1 - z_{i1}))$; and

$$g(\gamma, \tau | \ldots) \propto \prod_{i=1}^N \left(\frac{\exp\{\gamma z_{i1} + \tau x_i z_{i1}\}}{1 + \exp\{\gamma + \tau x_i\}}\right) \exp\left\{-\frac{\gamma^2 + \tau^2}{2{\sigma_0}^2}\right\}. \tag{6.8}$$

The variates γ and τ are generated by the Metropolis algorithm with (6.8) as the target distribution.

In addition to the normal distribution for the latency distribution of the subgroup 1, we also consider the log-normal distribution. Moreover, we consider probit links as opposed to logistic links for the mixing probabilities. We would like to answer the question: "Do we really need the mixture model? Would a simpler model with only one population suffice?" Therefore, we also fit the data with just the normal or the log-normal model. These models are the usual linear or log-linear models with Gaussian errors. We summarize the six models in Table 3.

On implementing for Model 2, we only need to change the original data to the log scale, and change the step in (6.7) to

$$\lambda|\ldots \sim \Gamma\left(N - z_{+1} + 1,\ \frac{1}{\sum_{i=1}^{N} e^{w_i}(1 - z_{i1})}\right).$$

For Models 3 and 4, we assume $p_1(\mathbf{x}, \boldsymbol{\rho}) = \Phi(\gamma + \tau x)$, and $p_2(\mathbf{x}, \boldsymbol{\rho}) = 1 - p_1(\mathbf{x}, \boldsymbol{\rho})$, where Φ is the standard cumulative normal distribution function. We follow the same procedures as before, except (6.8) is replaced by

$$g(\gamma, \tau|\ldots) \propto \prod_{i=1}^{N} \Phi(\gamma + \tau x_i)^{z_{i1}} (1 - \Phi(\gamma + \tau x_i))^{z_{i2}} \exp\left\{-\frac{\gamma^2 + \tau^2}{2{\sigma_0}^2}\right\}.$$

Then γ and τ can either be generated by the Metropolis algorithm or by using another data augmentation with two classes of latent normal variates; one is right-truncated at $\gamma + \tau x_i$, the other is left-truncated at the same point. For Models 5 and 6, we follow steps 1 and 3 of the Gibbs sampler in Section 4, where step 1 depends on a single model. Step 2 is not needed because we have a single linear or log-linear model. Then step 3 follows the usual Bayesian linear model updating.

6.3 Numerical Results

Most of our results are based on the male data given in Table 1. Table 2 lists the point estimates of the parameters computed from the EM algorithm and the MCMC algorithm for the model 1 given by (6.3). The 90% and 95% highest posterior density (HPD) intervals are computed from the Gibbs sampler. The Gibbs sampler is iterated 10,000 times and the results of the last 5,000 iterations are kept as a final sample from the joint posterior distribution. The incidence probabilities $p_1(x_c, \gamma, \tau) = \exp\{\gamma + \tau x_c\}/(1 + \exp\{\gamma + \tau x_c\})$ for the susceptible group are estimated by the Gibbs sampler at .266, .623, .872, .967 for the four dosages. The results are quite comparable to (.299, .724, .870, .960)=(43/144,50/69,47/54,48/50), the crude estimates of the proportion of beetles that react to the treatments (killed before day 13). Observe the 90% HPD interval of β includes zero; that is, there are no appreciable dose effects on survival times for the susceptible group. This finding is consistent with previous studies by Hewlett (1974) and Pack and Morgan (1990). However, dosage does affect mixing probabilities upon observing the point and the interval estimates of τ. Moreover, the mean survival times for the male beetles for each of the

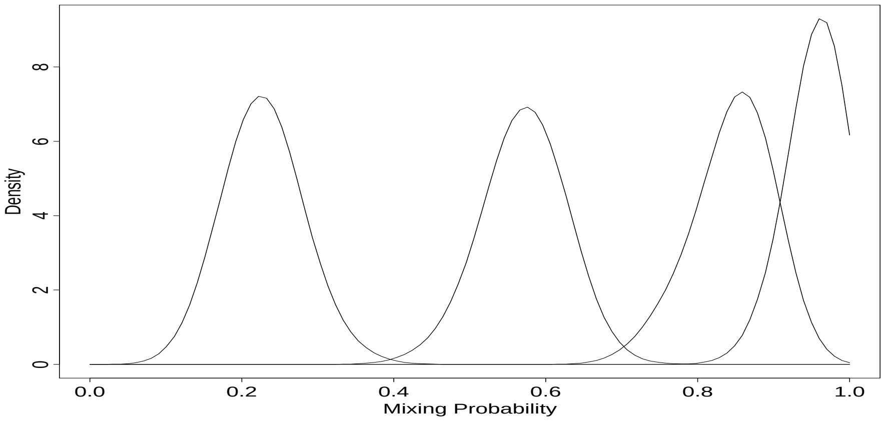

FIGURE 1. Plot of the posterior densities of the incidence probability $p_1(x_c, \gamma, \tau)$ for the subpopulation 1 evaluated at the four dosages. The plot is the overlay of the four histograms for the incidence probabilities for the susceptible group each evaluated at the log-dose level -1.61, -1.14, -0.69 and -0.22 mg/cm^2, respectively.

four dosages $\mu(x_c) = p_1(x_c, \gamma, \tau)(\alpha + \beta x_c) + p_2(x_c, \gamma, \tau)/\lambda$ are estimated at 125.8, 66.0, 24.2, and 8.5 from the Gibbs sampler.

Table 2: Parameter Estimates for the Male Data.

Parameter	MLE (EM)	Post. Mean (MCMC)	90% HPD Int.	95% HPD Int.
α	2.937	2.982	(2.350, 3.425)	(2.002, 3.729)
β	0.254	0.198	(-0.462, 0.652)	(-0.673, 0.721)
σ^2	2.217	2.253	(1.559, 2.924)	(1.148, 3.256)
λ	0.006	0.006	(0.004, 0.008)	(0.003, 0.009)
γ	4.213	4.197	(3.146, 5.318)	(2.986, 5.729)
τ	3.241	3.238	(2.466, 4.051)	(2.237, 4.124)

Figure 1 overlays the four histograms of the posterior incidence probabilities $p_1(x_c, \gamma, \tau)$ evaluated at the four log-dosages. It is clear it is monotonically increasing in the doses. The mixture model gives more weight to the normal distribution than to the exponential distribution for higher dose levels. This is consistent with fewer long-term survivors being observed at higher dosages. Figure 2 overlays the four posterior density plots for $\mu(x_c)$ for the four dosages respectively. The four densities reading from left to right correspond to the mean survival times for the beetles in the fourth column to the first column. The figure shows that all densities are unimodal. Their modes are quite comparable to the means given above. The mean decreases as the dosage increases.

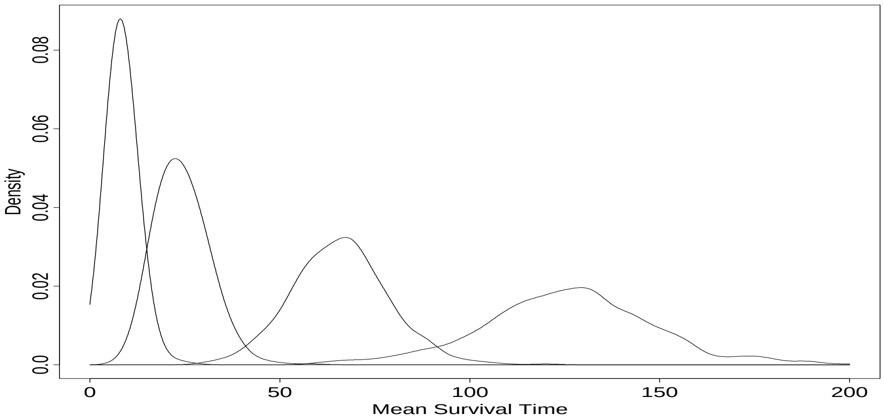

FIGURE 2. Plot of the posterior densities of the mean survival time $\mu(x_c)$ evaluated at the four dosages. The plot is the overlay of the four histograms (from left to right) for the mean survival times evaluated at the log-dose level -.22, -.69, -1.14, and -1.61 mg/cm^2 (in reverse order), respectively. Note the mean survival time increases as the dosage decreases.

We also analyzed the data set for the female beetles. The mean survival times for the females were estimated to be 163.3, 124.7, 74.4, and 34.0 for the four dosages. Apparently, there are big differences in the means between the sexes. The strong sex effect is also observed by Hewlett (1974), Diggle and Gratton (1984), and Pack and Morgan (1990). It is also interesting to note $\hat{\beta}$ is -.318, with (-.734, .099) as the 90% HPD interval. This shows that the dose has a stronger effect for the female than the male on the latency distribution for the susceptible group. As expected, the mean latent time of survival for the susceptible group decreases as dose increases.

For each of the six models, we evaluate the two-fold and four-fold cross-validated GCPOs. For the two-fold cross-validated GCPO, we randomly divide the data into two halves by randomly select the data points associated with half of the randomly chosen labels in 1–317. Similarly, we divide the data randomly into four groups of approximately equal sizes.

Table 3 lists the log product of the GCPO for each of the six models for the male data. The results show that the mixture models outperform the single-component model by a huge margin. Model 1 (6.3) is the best among the six models. Model 2 for a log-normal distribution is slightly inferior. The logistic link fits the data much better than the normal link. Finally, both two-fold and four-fold cross-validation yield similar results.

Table 3: Model Comparisons

Model	p_1	f_1	f_2	log PGCPO two-fold	log PGCPO four-fold
1	logistic	normal	exponential	-523.09	-522.73
2	logistic	log-normal	exponential	-524.09	-523.88
3	probit	normal	exponential	-526.66	-529.17
4	probit	log-normal	exponential	-527.33	-529.49
5	1	normal		-726.93	-773.79
6	1	log-normal		-587.18	-590.06

Table 4 gives the fitted values and the residuals for Model 1. To check the overall model fitting, we use the summation of absolute deviances, that is, $D_{abs} = \sum_{r=1}^{14}\sum_{c=1}^{4} |\text{observed} - \text{fitted}|$ for all the cells in Table 1. For model 1, this sum based on the first 13 days is 51.4, contrasting to 58.3 fitted by Pack and Morgan. In terms of this absolute deviance measure, model 1 also improves upon several models in Pack and Morgan.

Table 4: Fitted Values and Residuals

r/c	Fitted value				Residual			
	-1.61	-1.14	-0.69	-0.22	-1.61	-1.14	-0.69	-0.22
1	5.8	5.4	5.3	5.0	2.8	1.6	0.3	1.0
2	8.0	8.0	8.2	7.8	3.0	2.0	0.2	2.2
3	10.5	11.2	12.0	12.1	0.5	0.2	1.0	4.1
4	9.0	10.0	11.2	11.8	2.0	6.0	3.8	2.2
5	5.3	5.8	6.6	7.3	1.3	2.8	2.6	0.7
6	2.4	2.3	2.6	2.9	0.6	0.3	0.6	0.9
7	1.1	0.8	0.8	0.8	0.9	0.2	0.2	0.2
8	0.8	0.3	0.2	0.2	0.2	0.3	0.8	0.2
9	0.7	0.3	0.1	0.1	0.7	0.3	0.1	0.1
10	0.7	0.2	0.1	0.1	0.7	0.2	0.1	0.9
11	0.7	0.2	0.1	0.1	0.7	0.2	0.1	0.1
12	0.6	0.2	0.1	0.1	0.4	0.2	0.1	0.1
13	0.6	0.2	0.1	0.0	0.4	0.2	0.1	0.0
14	97.8	24.2	6.5	1.7	3.2	5.2	0.5	0.3

References

D. D. Boos and C. Brownie, "Mixture models for continuous data in dose-response studies when some animals are unaffected by treatment," *Biometrics* vol. 47, pp. 1489-1504, 1991.

D.R. Cox and D. Oakes, *Analysis of Survival Data,* Chapman and Hall: London, 1984.

A. Dempster, N. Laird, N and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society* Series B vol. 39, pp.1-38, 1977.

L. Devroye, *Non-Uniform Random Variate Generation,* Springer-Verlag: New York, 1986.

D. Dey, L. Kuo, and S. Sahu, "A Bayesian predictive approach to determining the number of components in a mixture distribution," *Statistics and Computing* vol. 5, pp. 297-305, 1995.

J. Diebolt and C. Robert, "Estimation of finite mixture distributions through Bayesian sampling," *Journal of the Royal Statistical Society* Series B vol. 56, pp. 363-375, 1994.

P.J. Diggle and R.J. Gratton, "Monte Carlo methods of inference for implicit statistical models," *Journal of the Royal Statistical Society* Series B vol. 46, pp. 193-227, 1984.

B. Efron and R. Tibshirani, *An Introduction to the Bootstrap,* Chapman & Hall: New York, 1993.

V.T. Farewell, "The use of mixture models for the analysis of survival data with long-term survivors," *Biometrics* vol. 38, pp. 1041-1046, 1982.

A. E. Gelfand, D. K. Dey, and H. Chang, "Model determination using predictive distributions with implementation via sampling-based methods (with Discussion)," In *Bayesian Statistics 4,* J. M. Bernardo, J. O. Berger, A.P. Dawid, and A. F. M. Smith (eds), pp. 147-169, Oxford University Press: Oxford, 1992.

A.E. Gelfand and A. F. M. Smith, "Sampling based approaches to calculating marginal densities," *Journal of the American Statistical Association* vol. 85, pp. 398-409, 1990.

A.E. Gelfand, A.F.M. Smith, and T.M. Lee, "Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling," *Journal of the American Statistical Association* vol. 87, pp. 523-532, 1992.

A. Gelman, and D. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical Science* vol. 7, pp. 457-472, 1992.

S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 6, pp. 721-741, 1984.

P. S. Hewlett, "Time from dosage to death in beetles Tribolium castaneum, treated with pyrethrins or DDT, and its bearing on dose-mortality relations," *Journal of Stored Product Research* vol. 10, pp. 27-41, 1974.

M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation* vol. 6, pp. 181-214, 1994.

M. I. Jordan, and L. Xu, "Convergence results for the EM approach to mixtures of experts architectures," Technical Report A. I. 1458, Massachusetts Institute of Technology, 1993.

A.Y.C. Kuk and C.H. Chen, "A mixture model combining logistic regression with proportional hazards regression," *Biometrika* vol. 79, pp. 531-541, 1992.

L. Kuo and A. F. M. Smith, "Bayesian computations in survival models via the Gibbs samplers (with Discussion)," In *Survival Analysis: State of the Art,* J. P. Klein and P. K. Goel (eds), pp. 11-14. Kluwer Academic: Dordrecht, 1992.

M. G. Larson and G. Dinse, "A mixture model for the regression analysis of competing risks data," *Applied Statistics* vol. 34, pp. 201-211, 1985.

P. McCullagh and J. A. Nelder, *Generalized Linear Models.* Chapman and Hall: London, 1989.

G. J. McLachlan and P. N. Jones, "Fitting mixtures models to grouped and truncated data via the EM algorithm," *Biometrics* vol. 44, pp. 571-578, 1988.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *Journal of Chemical Physics* vol. 21, pp. 1087-1092, 1953.

S. E. Pack and B. J. T. Morgan, B.J.T., "A mixture model for interval-censored time-to-response quantal assay data," *Biometrics* vol. 46, pp. 749-757, 1990.

F. Peng, R.A. Jacobs, and M. A. Tanner, "Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition," *Journal of the American Statistical Association* vol. 91, pp. 953-960, 1996.

M. A. Tanner and W. H. Wong, "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association* vol 82, pp. 528-550, 1987.

J. Taylor, "Semi-parametric estimation in failure time mixture models," *Biometrics* vol. 51, pp. 899-907, 1995.

C. G. Wei and M. Tanner, "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm," *Journal of the American Statistical Association*, 85, pp. 699-714, 1990.